

{7,"","","<I>","</I>","<TABLE>"}
procedure pcwrap HLRT (page <i>P</i>) scan to the line 7 in <i>P</i> while the next occurrence of <i>li</i> in <i>P</i> occurs before the next occurrence of <i>t</i> scan in <i>P</i> to next occurrence of <i>li</i> ; save position as start of item attribute scan in <i>P</i> to next occurrence of <i>ri</i> ; save position as end of item attribute scan in <i>P</i> to next occurrence of <i>lp</i> ; save position as start of price attribute scan in <i>P</i> to next occurrence of <i>rp</i> ; save position as end of price attribute return extracted, (item,price),. . . .} pairs

FIGURE 1
(PRIOR ART)

FIG. 1 is a flowchart illustrating a prior art procedure for extracting item and price attributes from a document. The procedure starts at a start node, proceeds to a decision node to check if the next occurrence of *li* in *P* occurs before the next occurrence of *t*. If yes, it proceeds to a scan node to scan in *P* to the next occurrence of *li*, save the position as the start of the item attribute, then scan in *P* to the next occurrence of *ri*, save the position as the end of the item attribute, then scan in *P* to the next occurrence of *lp*, save the position as the start of the price attribute, then scan in *P* to the next occurrence of *rp*, save the position as the end of the price attribute. If no, it proceeds to a return node to return the extracted, (item,price),. . . .} pairs. The procedure ends at an end node.

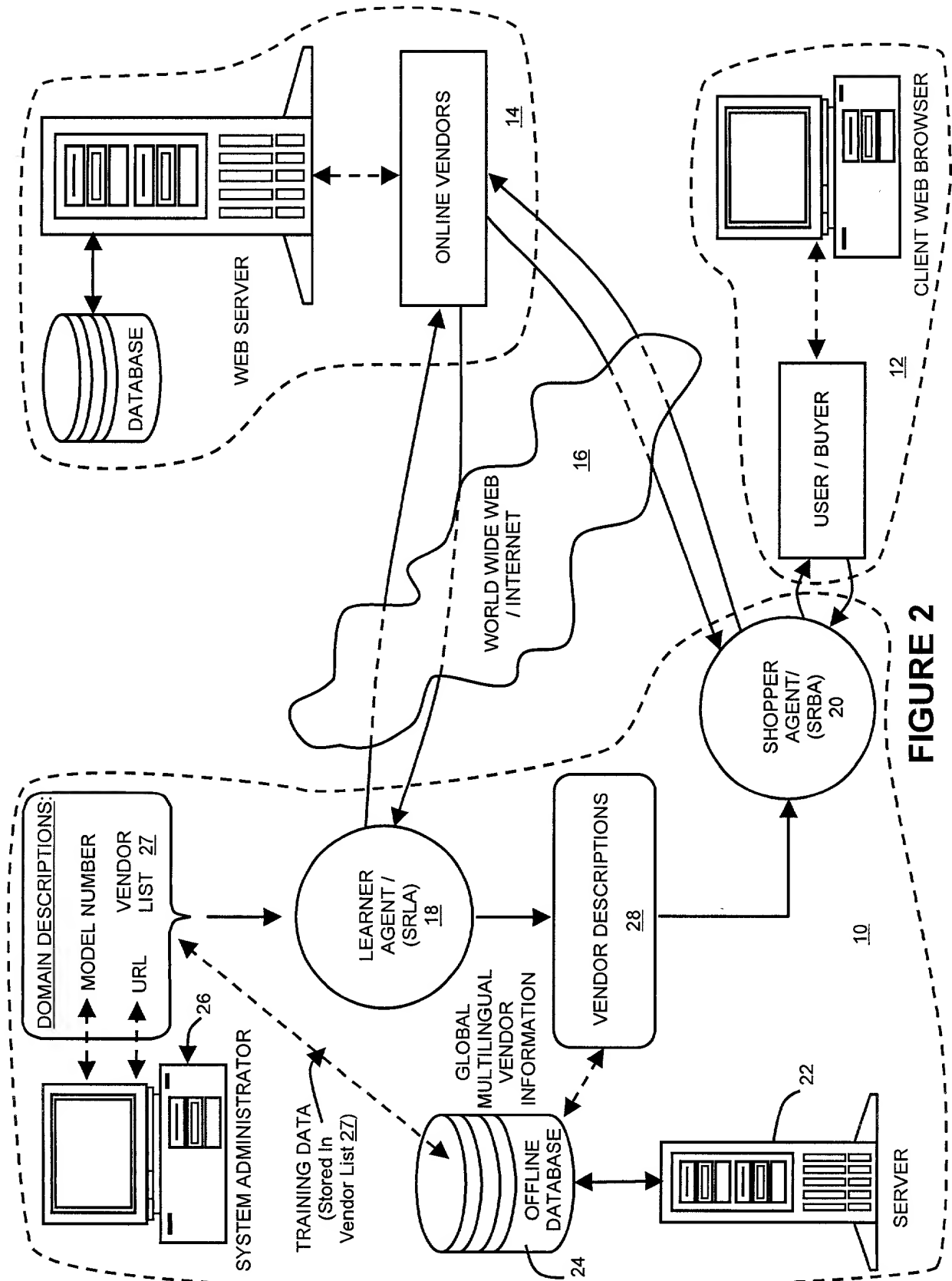


FIGURE 2

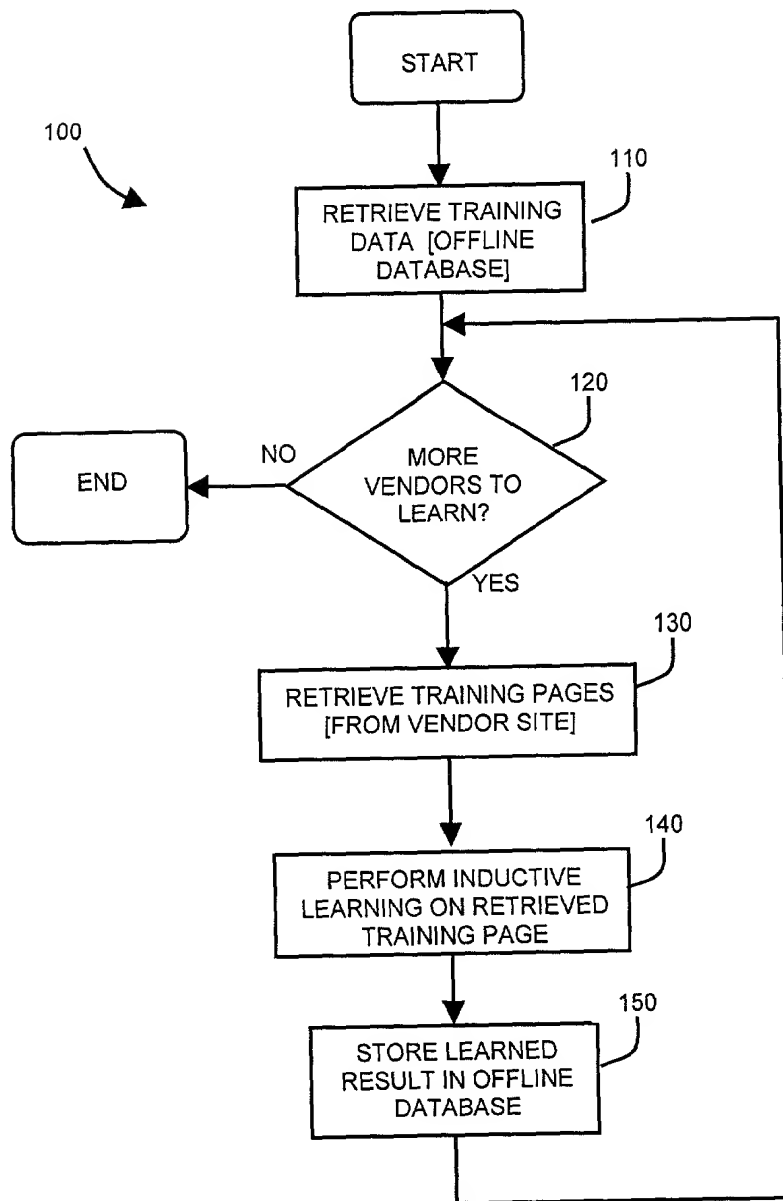


FIGURE 3

Element Name	Element Description
Vendor Name	Name of the online shop
Vendor URL	The URL of the online shop
Form URL	The URL to submit search data
Learning Domain	The domain used to learn the vendor description
Head	The end of header position of vendor's product pages. This element to learnt to reduce the product page searching space and thus shorten the product information extraction time
Tail	The start of footer position of vendor's product pages. Same as Head element, with this element the product information extraction time for Shopper Agent is shorten.
Left Delimiter of Item	The Shopper Agent uses this two delimiters to identify a product
Right Delimiter of Item	
Left Delimiter of Price	The Shopper Agent uses this two delimiters to locate a product's price information
Right Delimiter of Price	

FIGURE 4

EXAMPLE	
Element Name	Element Description
Vendor Name	800.com
Vendor URL	http://www.800.com
Form URL	http://www.800.com/search/srchrslts.asp?qs=1&sleentry=All&entry=
Learning Domain	Md
Head	1230
Tail	</BODY>
Left Delimiter of Item	/td width="12">
Right Delimiter of Item	</fon
Left Delimiter of Price	Your Price: \$
Right Delimiter of Price	

FIGURE 5

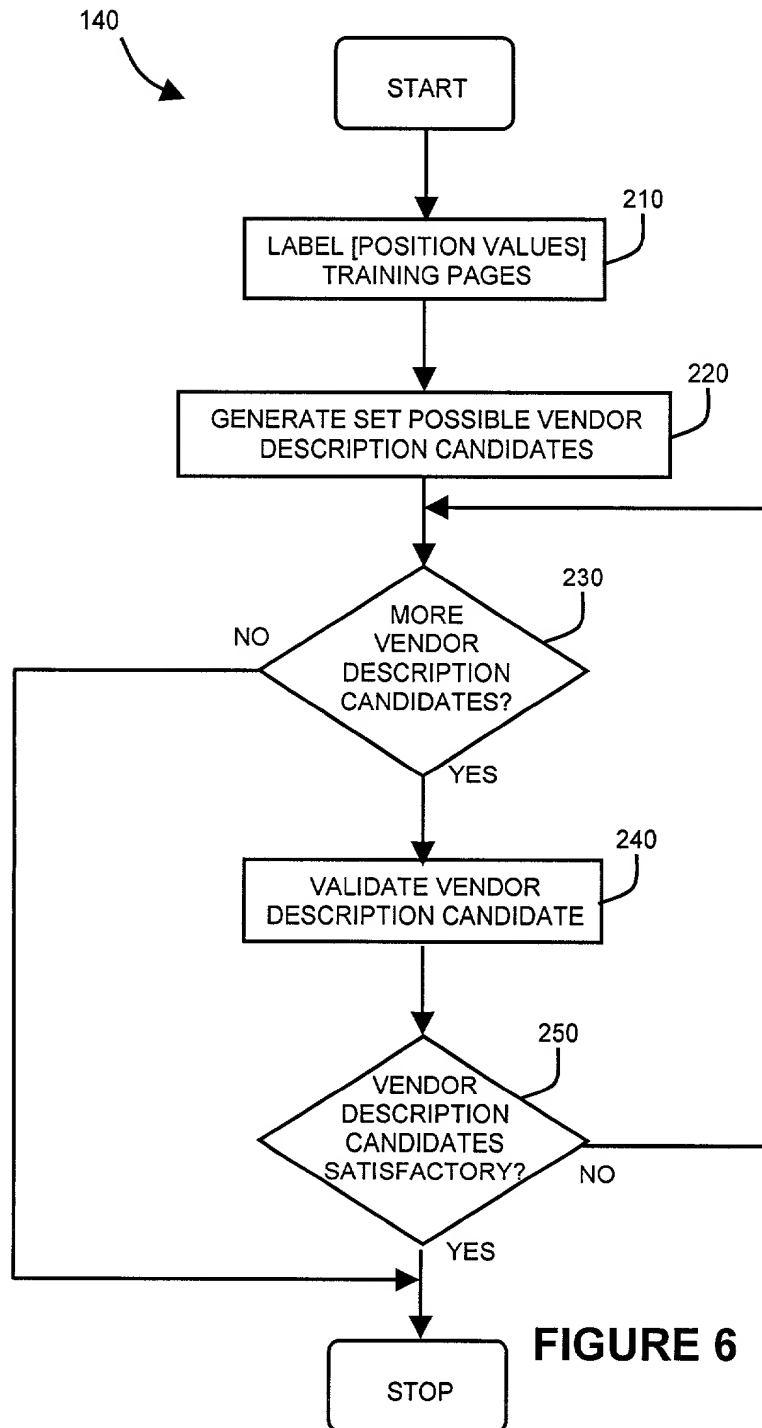


FIGURE 6

MD PRICE		<HTML> <TITLE>A Simple Product Catalogs</TITLE> <BODY> <H2> MD Price </H2> <TABLE BORDER=1> <TR BGCOLOR=ORANGE><TH> Model Number </TH> <TH> PRICE(US\$)</TH></TR> <TR><TD> HM381MD</TD><TD><I>399.95</I></TD></TR> <TR><TD> MD2070</TD><TD><I>599.95</I></TD></TR> <TR><TD> MD203</TD><TD><I>249.95</I></TD></TR> <TR><TD> MDR3</TD><TD><I>399.95</I></TD></TR> </TABLE> <P> <HR WIDTH=200 ALIGN=LEFT> <P> End Of The Product Catalog </BODY> </HTML>
Model Number	PRICE (US\$)	
HM381MD	399.95	
MD2070	599.95	
MD203	249.95	
MDR3	399.95	

FIGURE 7

Product Entry	Labels
{ HM381MD, 399.95 }	{<<174, 180>, <197, 202>>}
{ MD2070, 599.95 }	{<<229, 234>, <251, 256>>}
{ MD203, 249.95 }	{<<283, 287>, <304, 309>>}
{ MDR3, 399.95 }	{<<336, 339>, <356, 361>>}

FIGURE 8

Product Entry	Labels
{ PRODUCT, PRICE }	{<PRODUCT LEFT DELIMITER, PRODUCT RIGHT DELIMITER>, <PRICE LEFT DELIMITER, PRICE RIGHT DELIMITER>}
***	***

FIGURE 9

Delimiter	Candidates
Item Left Delimiter	<TR><TD> TR><TD> R><TD> ><TD> <TD> TD> D> > B> >
Item Right Delimiter	</TD><TD>< > </TD><TD>< </TD><TD>< </TD><TD> </TD><TD </TD><T </TD>< </TD> </TD </T </ < </B </ <
Price Left Delimiter	</TD><TD>< > /B></TD><TD>< > B></TD><TD>< > ></TD><TD>< > </TD><TD>< > /TD><TD>< > TD><TD>< > D><TD>< > ><TD>< > <TD>< > TD>< > D>< > >< > < > > >

FIGURE 10A

[illegible]

FIGURE 10B

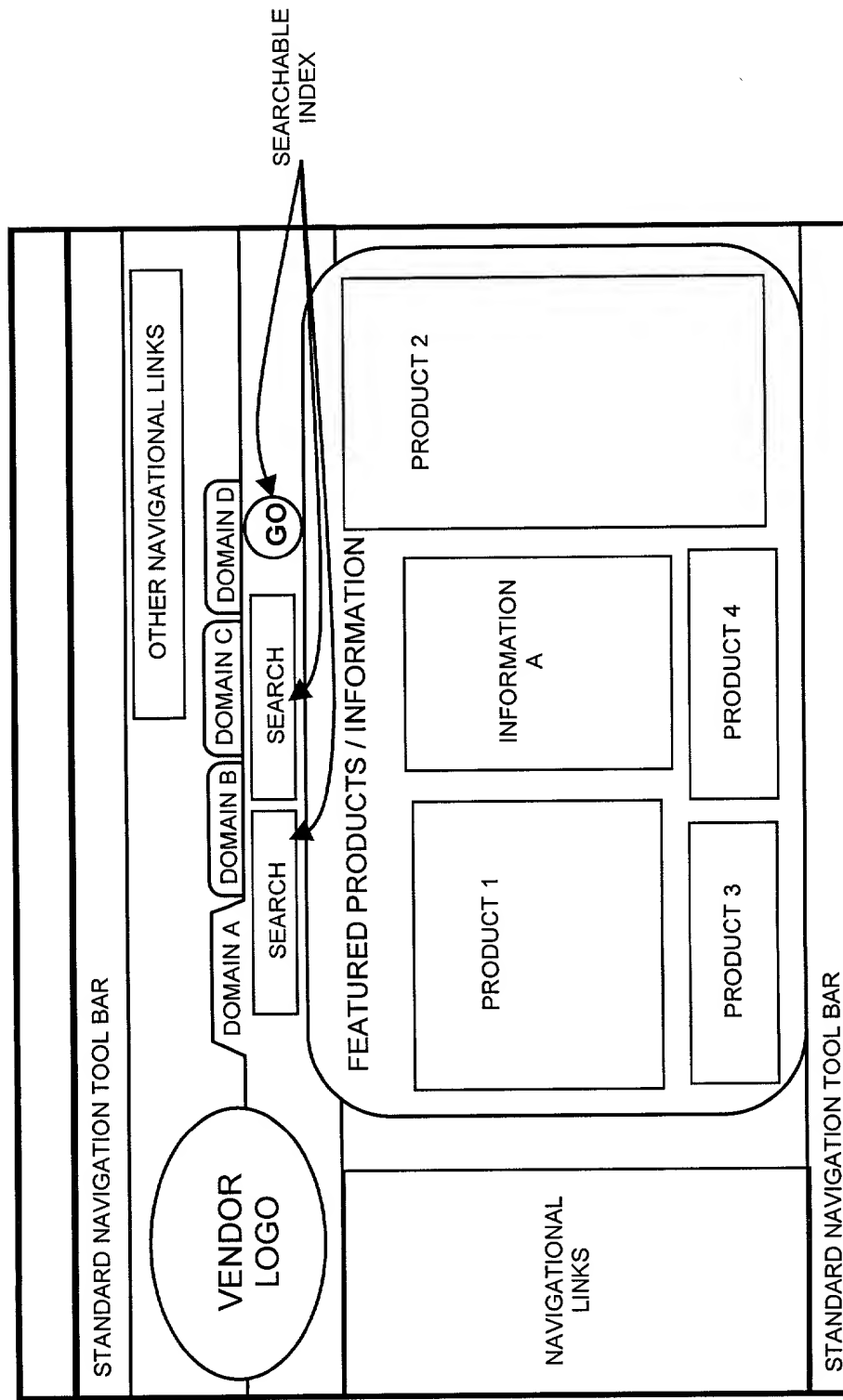


FIGURE 11

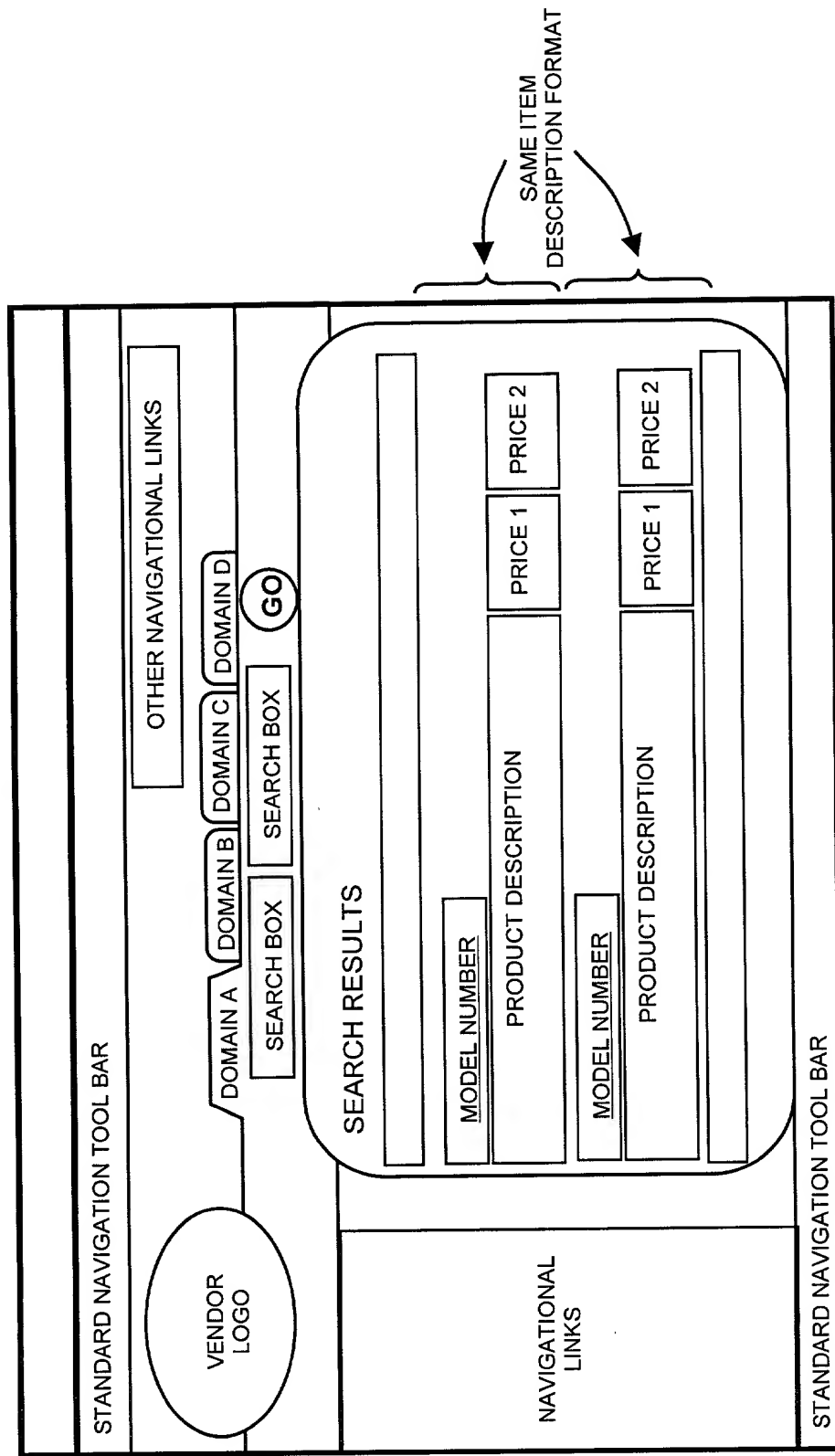


FIGURE 12

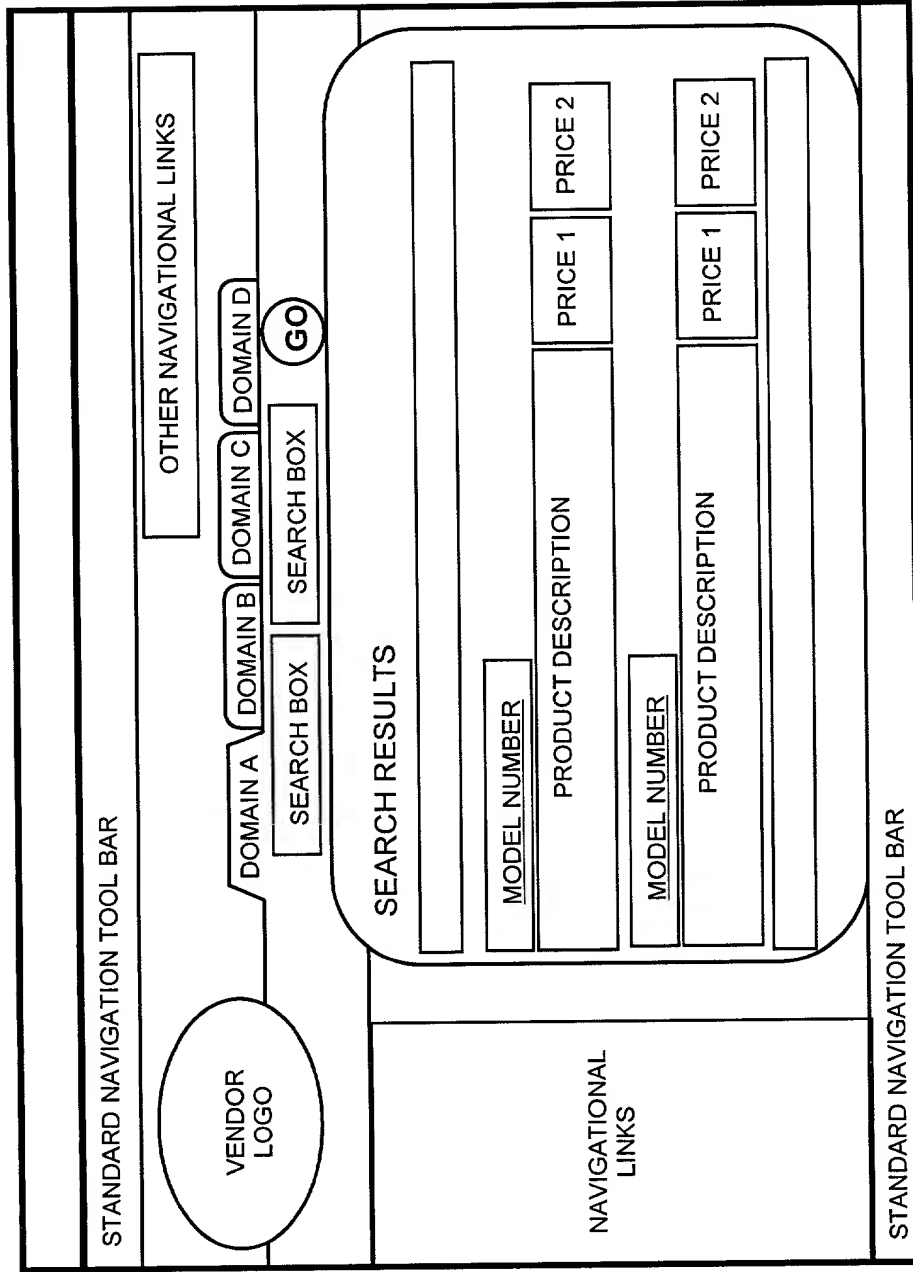


FIGURE 13

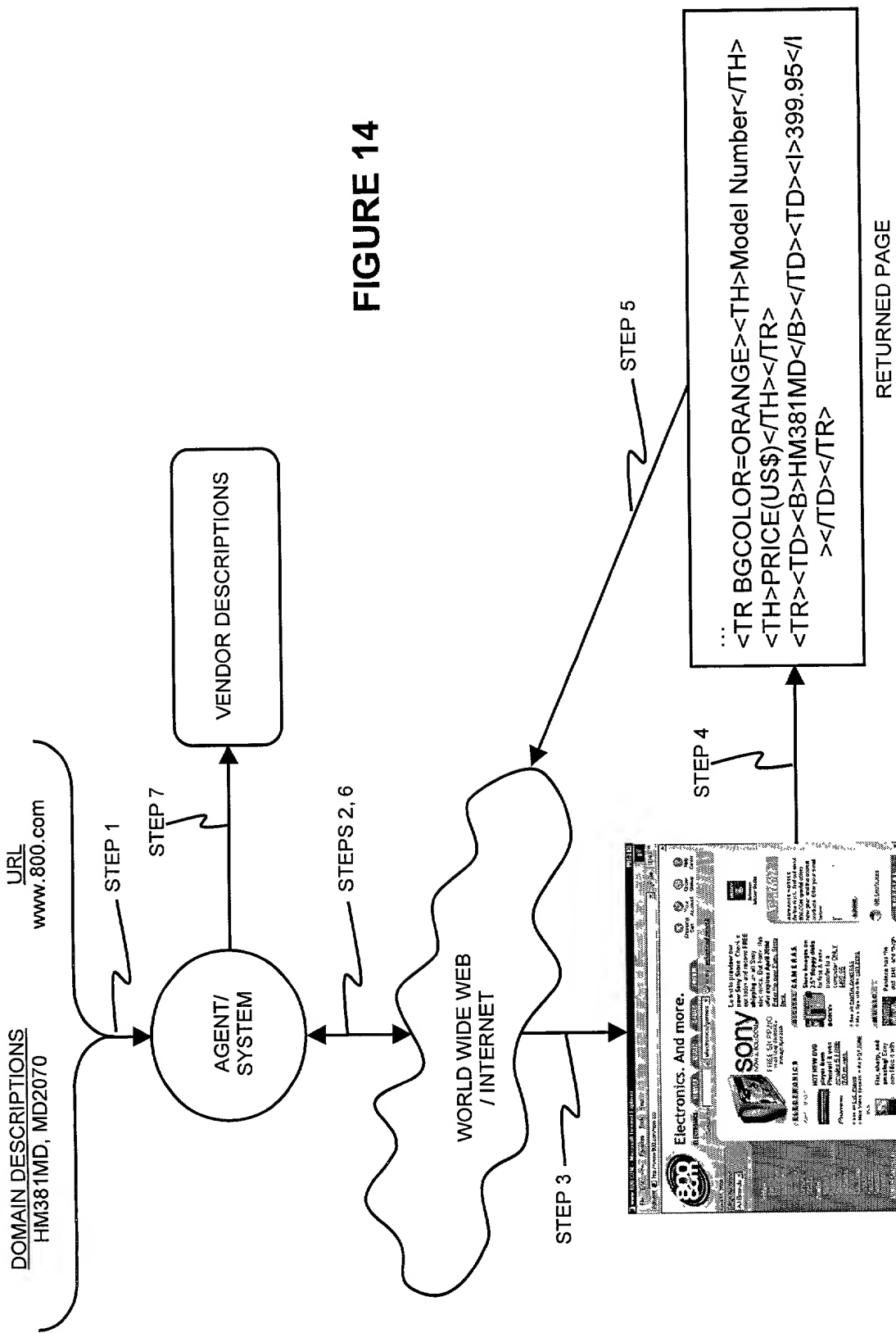


FIGURE 14

RETURNED PAGE

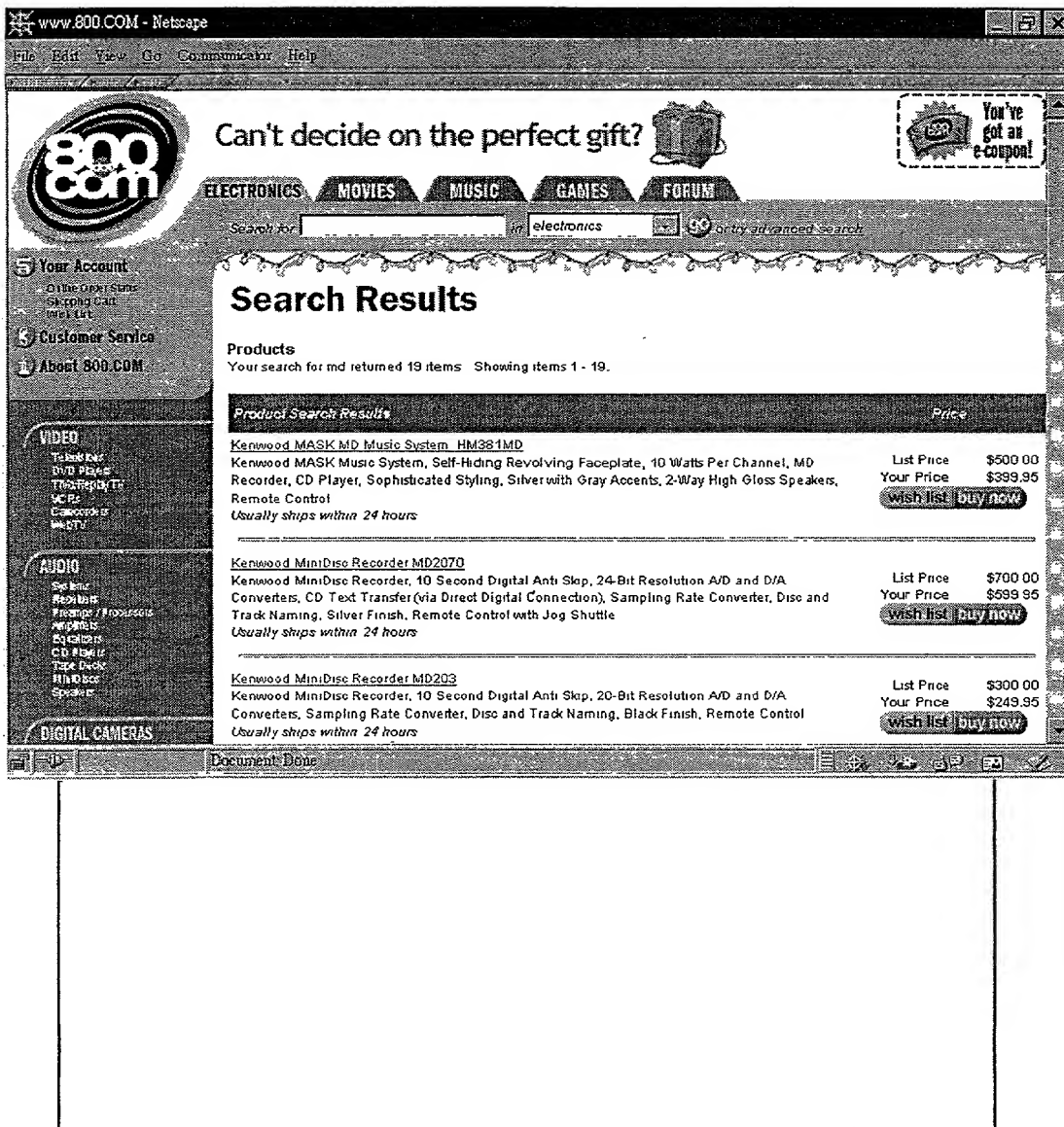


FIGURE 15A

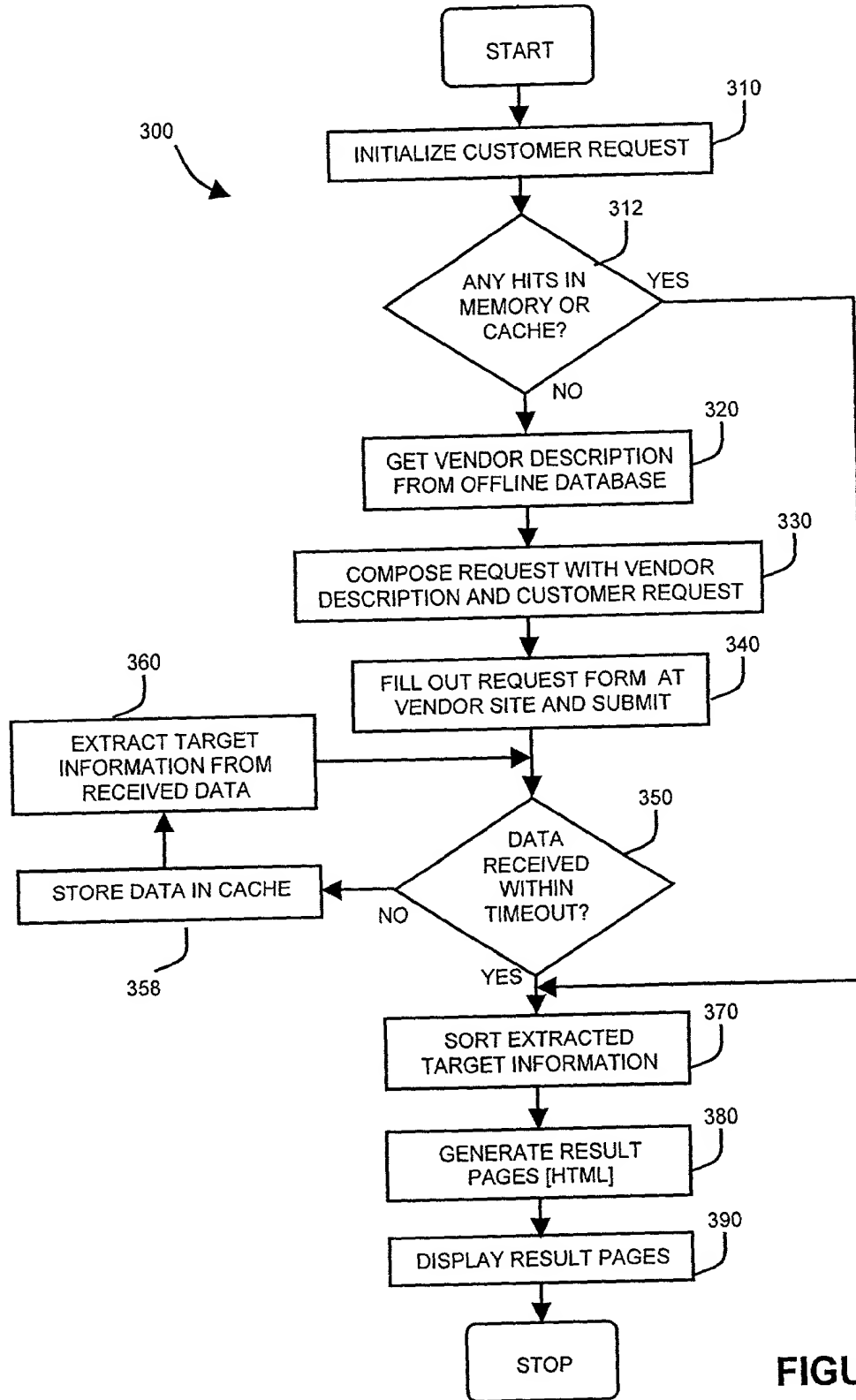


FIGURE 15C

Vendor Information [X]

Vendor Information | Add Vendor

Add Vendor

Details

☒ Provide Training Examples

Vendor Name:

URL:

Form's URL:

Learning Domain:

☐ Enter Wrapper For This Vendor

Head:

Tail:

Left Delimiter of Item:

Right Delimiter of Item:

Left Delimiter of Price:

Right Delimiter of Price:

Save

OK Cancel Apply

FIGURE 16

Vendor Information [X]

Vendor Information: Add Vendor

Add Vendor

Details

Vendor Name:

URL:

Form's URL:

Learning Domain:

☒ Provide Training Examples

☐ Enter Wrapper For This Vendor

Head:

Tail:

Left Delimiter of Item:

Right Delimiter of Item:

Left Delimiter of Price:

Right Delimiter of Price:

Save

OK Cancel Apply

FIGURE 17

Vendor Information

Vendor Information | Add Vendor

Vendor Name: 1cache.com Search

Details

Vendor Name: 1cache.com

URL: http://www.1cache.com

Form's URL: http://st4.yahoo.com/cgi-bin/rssearch?catalog=1ca

Learning Domain: dvd

Wrapper

Head: Tail:

Left Delimiter of Item: Right Delimiter of Item:

Left Delimiter of Price: Right Delimiter of Price:

Save

OK Cancel Apply

FIGURE 18

Vendor Information

Vendor Information | Add Vendor

Vendor Name: Search

Details:

Vendor Name:

URL:

Form's URL:

Learning Domain:

Wrapper:

Head:

Tail:

Left Delimiter of Item:

Right Delimiter of Item:

Left Delimiter of Price:

Right Delimiter of Price:

Save

OK Cancel Apply

FIGURE 19

Training Data

Training Example

Vendor:

Vendor	Training Example
1cache.com	DVD Virtual Notebook Theater with I-Glas
1cache.com	JVC XV523GD Dolby Digital DVD Player
1cache.com	Pioneer DVL-919 - Combination LD/DVD

Vendor:
Example:

FIGURE 20

Learn Vendor Description

Learn Vendor Description

Learning Option

☒ Learn All
☐ Learn One

Vendor Name:

FIGURE 21

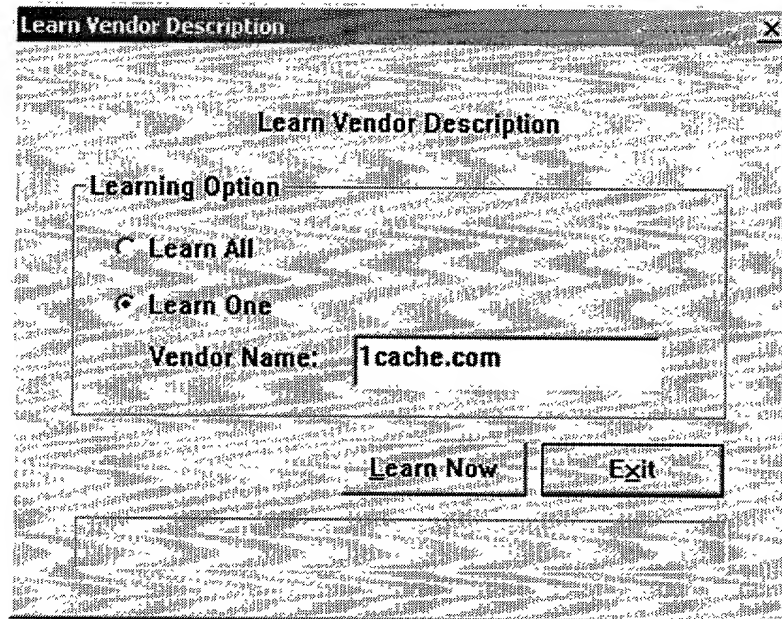


FIGURE 22

Vendor Information [X]

Vendor Information [Add Vendor]

Vendor Information

Vendor Name: 1cache.com [Search]

Details:

Vendor Name: 1cache.com

URL: http://www.1cache.com

Form's URL: http://st4.yahoo.com/cgi-bin/nsearch?catalog=1ca

Learning Domain: dvd

Wrapper

Head: 5230

Tail: D></TD><TD></TD></TR></

Left Delimiter of Item: G SRC=/img/trans_1x1.gif BORDER=0 WIC

Right Delimiter of Item:

Left Delimiter of Price: </TD> <TD ALIGN=right><FON

Right Delimiter of Price: </T

[Save]

OK Cancel Apply

FIGURE 23

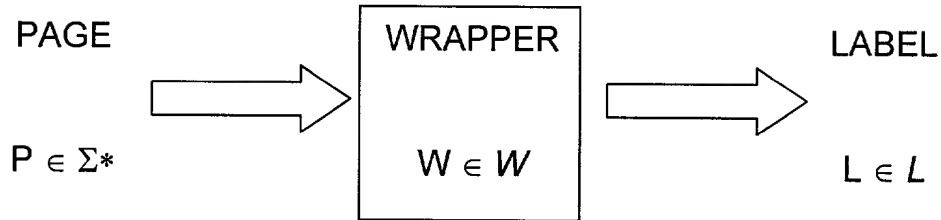


FIGURE 24

```

    procedure execHLRT (wrapper (h, li, ri, lp, rp, t), page P)
        m ← 0
        scan in P to line h
        while the next occurrence of li in P occurs before the next occurrence of t
            m ← m + 1
            scan in P to the next occurrence of li in P; save position as bm,i
            scan in P the next occurrence of ri; save position as em,i
            scan in P to the next occurrence of lp in P; save position as bm,p
            scan in P to next occurrence of rp; save position as em,p
        return label { . . . , <<bmi, emi>, <bmp, emp>> }
    
```

FIGURE 25

```

    Procedure learnHLRT (examples ε)
    1. Generate the candidate sets Candsr(l, ε), Candsl(l, ε), Candsr(p, ε), Candsl(p, ε).
    2. Enumerate the cross product of these candidate sets; each element W = (h, li, ri, lp, rp, t), of this cross product is a wrapper. Halt if W is satisfactory i.e., execLR(W, Pn) = Ln, for every (Pn, Ln) ∈ ε.
    
```

FIGURE 26

```

/*
* The main problem
*/
procedure learnHLRT (examples  $\epsilon$ ) for each  $1 < k < K$ 
    for each  $u \in \text{cands}_l(p, \epsilon)$ : if  $\text{valid}_l(u, p, \epsilon)$  then  $l_p \leftarrow u$  and terminate this loop
    for each  $u \in \text{cands}_r(i, \epsilon)$ : if  $\text{valid}_r(u, i, \epsilon)$  then  $r_i \leftarrow u$  and terminate this loop
    for each  $u \in \text{cands}_l(p, \epsilon)$ : if  $\text{valid}_r(u, p, \epsilon)$  then  $r_p \leftarrow u$  and terminate this loop
    for each  $u_{li} \in \text{cands}_l(i, \epsilon)$ 
    for each  $u_h \in \text{cands}_h(\epsilon)$ 
        for each  $u_t \in \text{cands}_t(\epsilon)$ 
            if  $\text{valid}_{li,ht}(u_{li}, u_h, u_t, \epsilon)$  then
                 $l_i \leftarrow u_{li}$ ,  $h \leftarrow u_h$ ,  $t \leftarrow u_t$ , and terminate these three loops
    return HLRT wrapper ( $h, l_i, r_i, l_p, r_p, t$ )

/*
* Generate a set of candidates for left delimiter of price attribute
*/
procedure  $\text{cands}_l(\text{attribute price, examples } \epsilon)$ 
    return the set of all suffixes of the shortest string in  $\text{neighbors}_l(\text{price}, \epsilon)$ 

/*
* Generate a set of candidates for right delimiter of attribute a. Here a could be item or price
*/
procedure  $\text{cands}_r(\text{attribute a, examples } \epsilon)$ 
    return the set of all prefixes of the shortest string in  $\text{neighbors}_r(a, \epsilon)$ 

/*
* Generate a set of candidates for a page's head
*/
procedure  $\text{cands}_h(\text{examples } \epsilon)$ 
    return the set of all substrings of the shortest string in  $\text{heads}(\epsilon)$ 

/*
* Generate a set of candidates for a page's tail
*/
procedure  $\text{cands}_t(\text{example } \epsilon)$ 
    return the set of all substrings of the shortest string in  $\text{tails}(\epsilon)$ 

/*
* Determine whether a particular candidate for the left delimiter of price is valid
* This procedure applies constraints C3
*/
procedure  $\text{valid}(\text{candidate } u, \text{attribute price, examples } \epsilon)$ 
    for each  $s \in \text{neighbors}_l(\text{price}, \epsilon)$ : if  $u$  is not a proper suffix of  $s$  then return FALSE
    
```

Figure 27A


```

    for each  $s \in \text{tails}(\epsilon)$ : if  $u$  is a sub string of  $s$  then return FALSE
    return TRUE

/*
 *Determine whether a particular candidate for the right delimiter of item or price is
 *valid
 *This procedure applies constraints C1 and C2
 */
    procedure validi(candidate  $u$ , attribute  $a$ , examples  $\epsilon$ )
    for each  $s \in \text{attrs}(a, \epsilon)$ : if  $u$  is a sub string of  $s$  then return FALSE
    for each  $s \in \text{neighbors}(a, \epsilon)$ : if  $u$  is not a prefix of  $s$  then return FALSE
    return TRUE

/*
 *Determine whether a particular combination of candidate  $u_h$ ,  $u_t$ , and  $u_l$  for  $h$ ,  $t$ , and
 *li respectively are satisfactory.
 */
    procedure validli,h,t(candidates  $u_{li}$ ,  $u_h$ ,  $u_t$ , example  $\epsilon$ )
    for each  $s \in \text{heads}(\epsilon)$ 
        if  $u_h$  is not a sub string of  $s$  then return FALSE
        if  $u_l$  is not a proper suffix of scan ( $s$ ,  $u_h$ ) then return FALSE
        if  $u_t$  occurs before  $u_l$  in scan ( $s$ ,  $u_h$ ), then return FALSE
    for each  $s \in \text{tails}(\epsilon)$ 
        if  $u_l$  is not a sub string of  $s$  then return FALSE
        if  $u_l$  occurs before  $u_t$  in  $s$  then return FALSE
    for each  $s \in \text{seps}(\epsilon)$ 
        if  $u_l$  is not a proper suffix of  $s$  then return FALSE
        if  $u_t$  occurs before  $u_l$  in  $s$  then return FALSE
    return TRUE

/*
 *Return a set of containing all values of a attribute, either item( $i$ ) or price( $p$ ) in each
 *example
 */
    procedure attrs(attribute  $a$ , examples  $\epsilon$ )
    return  $U_{(Pn,Ln)} \in \epsilon(Pn[b_{m,a}, e_{m,a}] | \langle \dots, (b_{m,k}, e_{m,k}), \dots \rangle \in Ln)$ 

/*
 *Return all strings to the left of an attribute, whether these strings are in the heads
 * or the bodies of the pages
 */
    procedure neighborsi(attribute  $a$ , examples  $\epsilon$ )
    if  $a=i$  then return seps ( $i, \epsilon$ )  $\cup$  heads ( $\epsilon$ ) else return seps ( $a, \epsilon$ )

```

FIGURE 27B

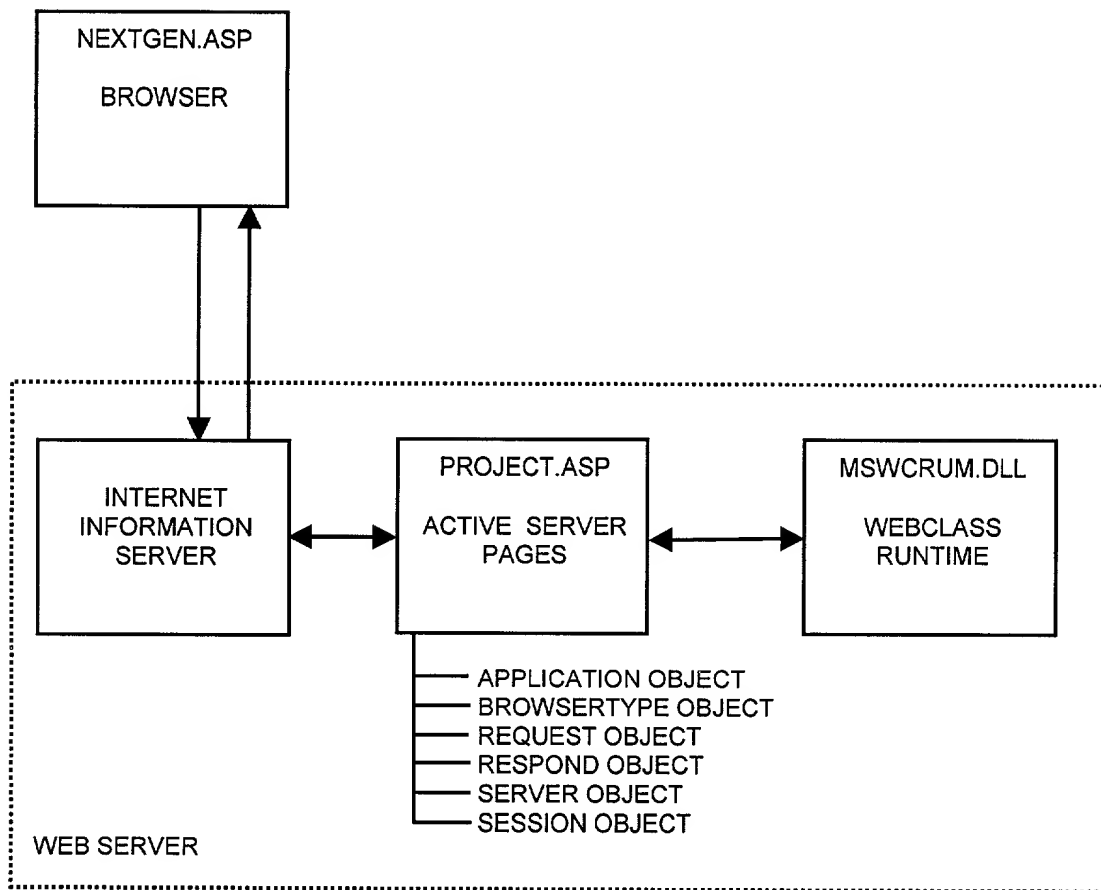


FIGURE 28

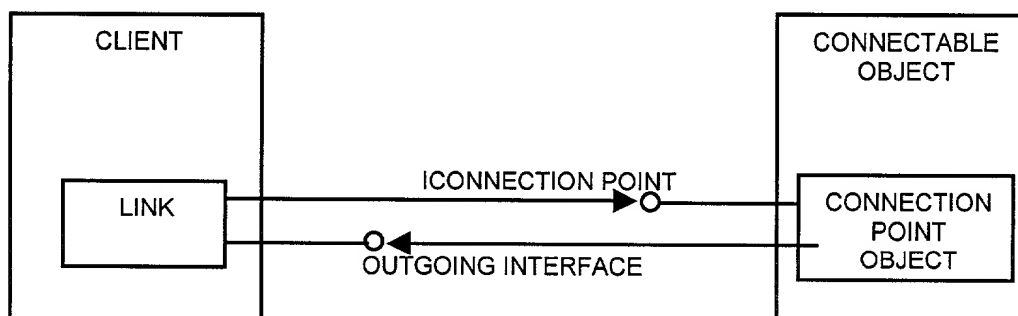


FIGURE 29

FIGURE 30

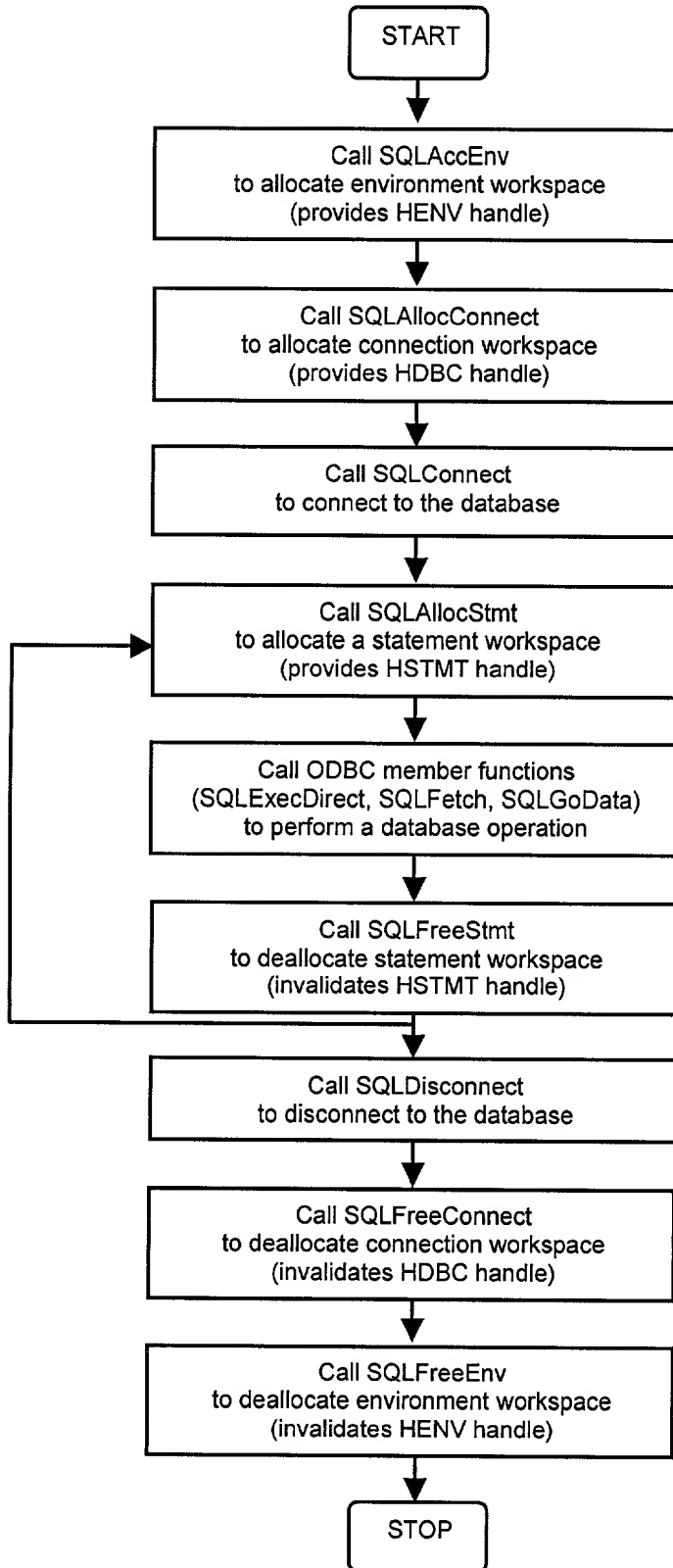
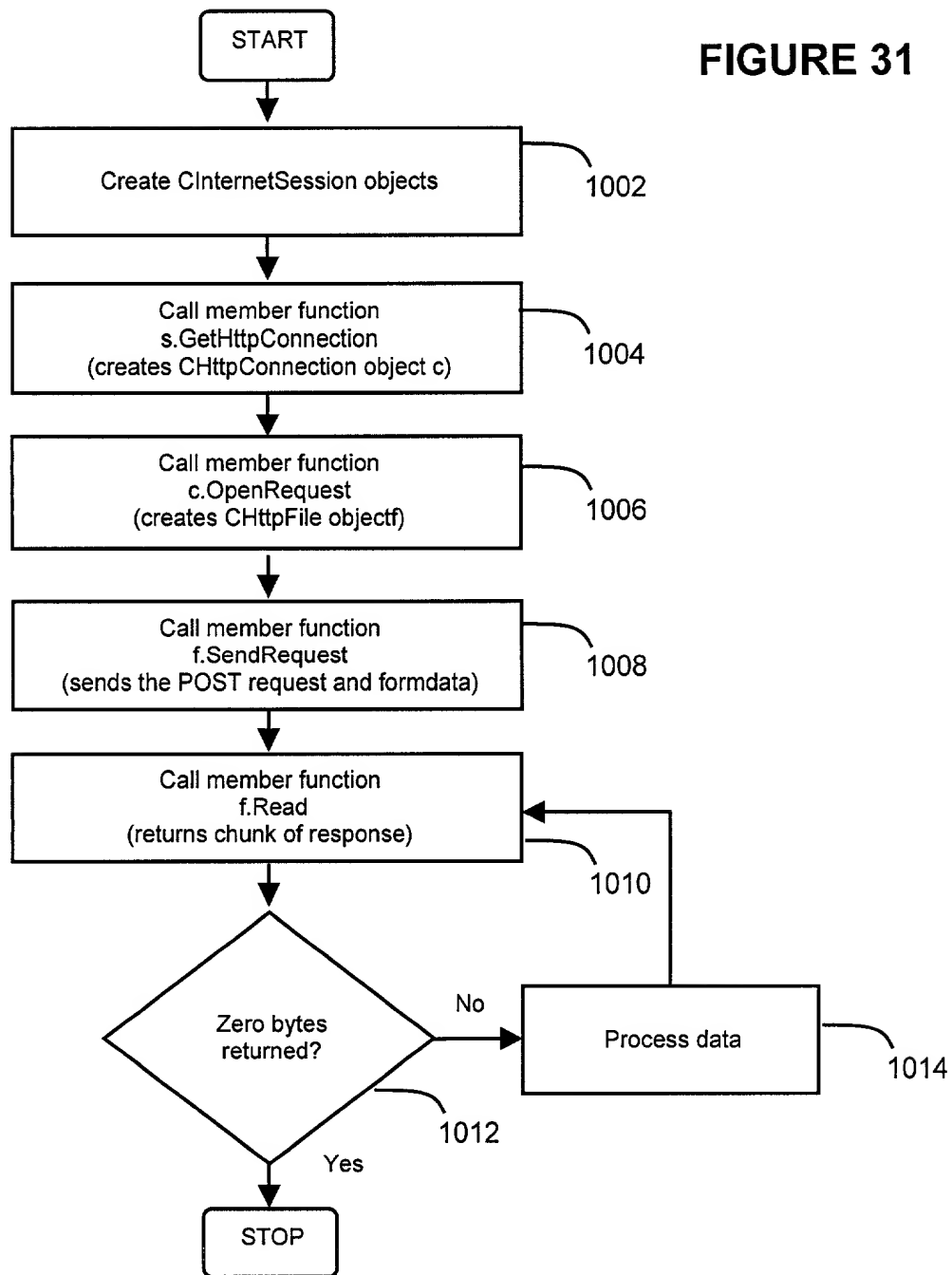


FIGURE 31



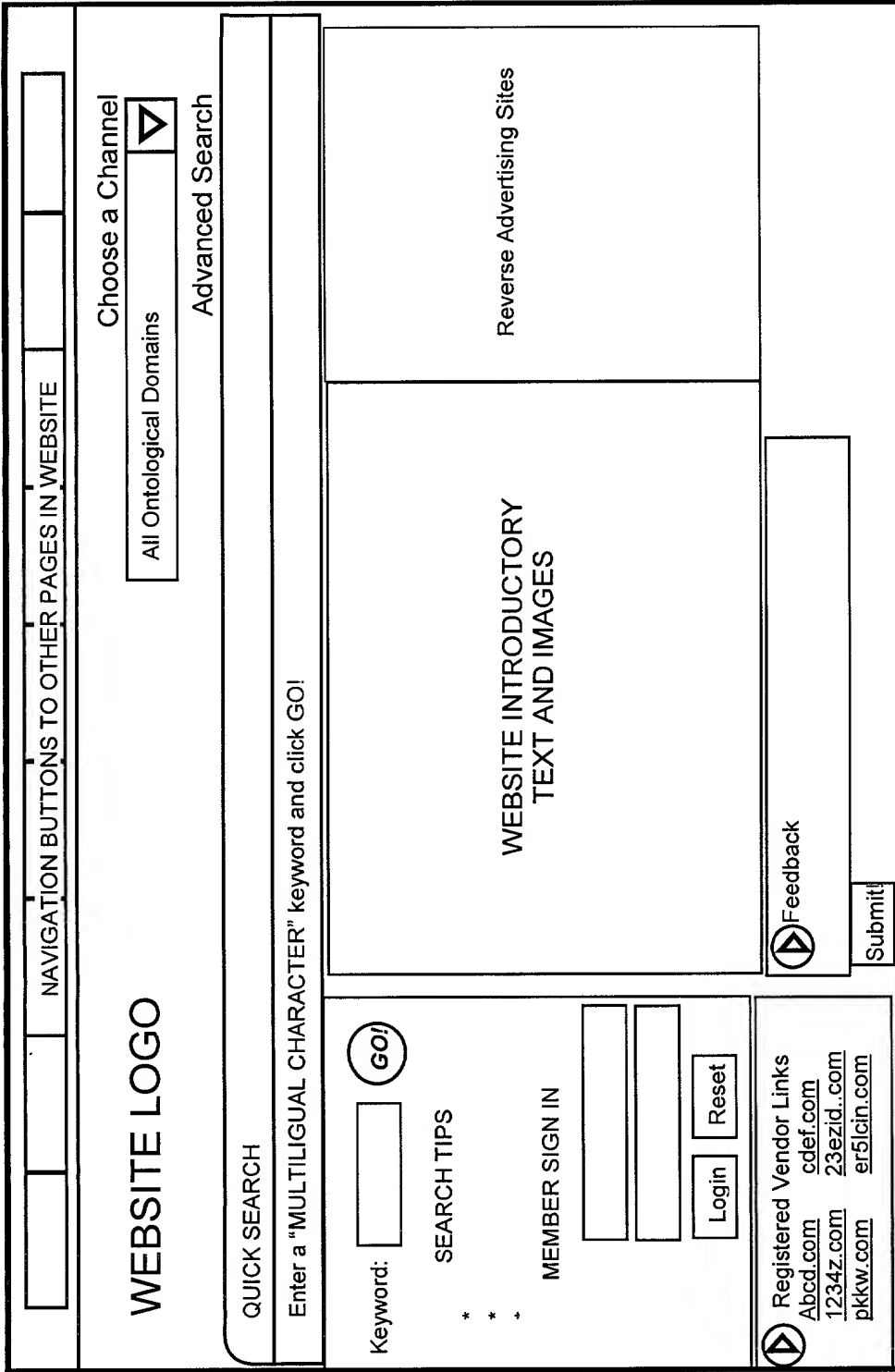


FIGURE 32

NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE

WEBSITE LOGO

Choose a Channel

ALL G2B PROCUREMENT DOMAINS

Advanced Search

DOMAIN A

DOMAIN B

DOMAIN C

DOMAIN D

OTHER DOMAINS

QUICK SEARCH

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

Keyword:

GO!

SEARCH TIPS

*
*
*

Company A

Information about G2B Company A

Company G

Information about G2B Company G

Reverse Advertising Sites

Registered Vendor Links

Abcd.com

1234z.com

pkkw.com

cdef.com

23ezid..com

er5icin.com

Feedback

Submit!

FIGURE 33

NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE

WEBSITE LOGO

Choose a Channel

All G2B Tender Procurement Domains

Advanced Search

DOMAIN A

DOMAIN B

DOMAIN C

DOMAIN D

OTHER DOMAINS

ADVANCED SEARCH AGENTS ARE ON!

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

SEARCH TIPS

☐ Select All

Public Body

Speed

Air/Shipping/Ground/Free Overnight Del.

Ease of Navigation

☐ G2B Company G

⇅

EEEE

Time Out

Reserved Price Range

Manufacturer or OEM

Keyword:

From \$

To \$

GO!

FIGURE 34

NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE

WEBSITE LOGO

Choose a Channel

All Ontological Domains

Advanced Search

DOMAIN A

DOMAIN B

DOMAIN C

DOMAIN D

OTHER DOMAINS

QUICK SEARCH

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

Keyword:

GO!

SEARCH TIPS

*
*
*

For B2B
Trading
Exchanges,
Auctions,
Vertical
Portals,
Production
and Non-
Production e-
Procurement
Supplies

Agricultural / Farming
Apparel
Automotive/ Transportation
Business Machines
Computers & Electronics
Industrial Equipment
Medical & Scientific
Equipment

More Business to
Business
MRO/Raw Materials
Office Furnishings
Office Supplies
Real Estate
Restaurant/Hospitality
Services
Travel & Entertainment

Reverse Advertising Sites

Registered Vendor Links
[Abcd.com](#)
[1234z.com](#)
[pkkw.com](#)

Feedback

Submit

Registered Vendor Links
[cdef.com](#)
[23ezid..com](#)
[er5lcin.com](#)

FIGURE 35

	NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE		Choose a Channel	
--	--	--	------------------	--

WEBSITE LOGO

All Ontological Domains ▼

Advanced Search

DOMAIN A	DOMAIN B	DOMAIN C	DOMAIN D	OTHER DOMAINS
ADVANCED SEARCH AGENTS ARE ON!				

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

<input type="checkbox"/> Select All	Multilingual B2B Suppliers			
Online Stores	Speed	Air/Shipping/Ground/Free Overnight Del.	Ease of Navigation	
<input type="checkbox"/> B2B Company 1	↑↑↑↑↑	⇒	EEEE	
<input type="checkbox"/> B2B Company 2	↑↑↑↑↑	⇒	EEEE	
<input type="checkbox"/> B2B Company 3	↑↑↑↑↑	⇒	EEEE	
<input type="checkbox"/> B2B Company 4	↑↑↑↑↑	⇒	EEEE	

SEARCH TIPS

* * *

Time Out

Price Range

From \$
To \$

Manufacturer

Keyword:

GO!

FIGURE 36

NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE

WEBSITE LOGO

Choose a Channel

All Domains

▼

Advanced Search

DOMAIN A

DOMAIN B

DOMAIN C

DOMAIN D

OTHER DOMAINS

QUICK SEARCH

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

Keyword:

GO!

SEARCH TIPS

*
*
*

Hot Sellers Domain A

Image Item A	Description Item A	Image Item B	Description Item B
Image Item 1	Description Item 1	Image Item G	Description Item G
Image Item A	Description Item A	Image Item K	Description Item K

Registered Vendor Links

Abcd.com

cdef.com

1234z.com

pkkw.com

Feedback

Submit!

FIGURE 37

NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE

WEBSITE LOGO

Choose a Channel

All Domains

Advanced Search

DOMAIN A

DOMAIN B

DOMAIN C

DOMAIN D

OTHER DOMAINS

ADVANCED SEARCH AGENTS ARE ON!

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

SEARCH TIPS

English Vendors

Chinese Vendors

Time Out

Price Range

Manufacturer

Keyword:

FIGURE 38

NAVIGATION BUTTONS TO OTHER PAGES IN WEBSITE

WEBSITE LOGO

Choose a Channel

ALL DOMAINS

Advanced Search

DOMAIN A

DOMAIN B

DOMAIN C

DOMAIN D

OTHER DOMAINS

QUICK SEARCH

Enter a "MULTILINGUAL CHARACTER" keyword and click GO!

Keyword:

GO!

SEARCH TIPS

Search Results

Products

Your search for XX returned 5 Domain C items.

Model / Features	Store	Price
<u>AMS7 XX Widget</u> [Features]	Store 1443A	\$399.95 Buy now
<u>AMS2 XX Widget</u> [Features]	Store 1443A	\$329.95 Buy now
<u>AMS7 XX Widget</u> [Features]	Store DRGWS4	\$459.95 Buy now
<u>AMS7 XX Widget</u> [Features]	Store 4U\$U\$	\$499.95 Buy now
<u>ZRTQ XX Widget</u> [Features]	Store OOHE	\$529.95 Buy now

FIGURE 39